Temporal similarity metrics for latent network reconstruction: The role of time-lag decay

Liao, Hao ; Liu, Ming-Kai ; Mariani, Manuel ; Zhou, Mingyang ; Wu, Xingtong

Abstract: When investigating the spreading of a piece of information or the diffusion of an innovation, we often lack information on the underlying propagation network. Reconstructing the hidden propagation paths based on the observed diffusion process is a challenging problem which has recently attracted attention from diverse research fields. To address this reconstruction problem, based on static similarity metrics commonly used in the link prediction literature, we introduce new node-node temporal similarity metrics. The new metrics take as input the time-series of multiple independent spreading processes, based on the hypothesis that two nodes are more likely to be connected if they were often infected at similar points in time. This hypothesis is implemented by introducing a time-lag function which penalizes distant infection times. We find that the choice of this time-lag function strongly affects the metrics' reconstruction accuracy, depending on the network's clustering coefficient, and we provide an extensive comparative analysis of static and temporal similarity metrics for network reconstruction. Our findings shed new light on the notion of similarity between pairs of nodes in complex networks.

DOI: https://doi.org/10.1016/j.ins.2019.01.081

Temporal similarity metrics for latent network reconstruction: The role of time-lag decay

Hao Liao[a], Ming-Kai Liu[a], Manuel Sebastian Mariani[b,c], Mingyang Zhou[a], Xingtong Wu[a]

[a] *National Engineering Laboratory on Big data Application on Improving Government Governance Capabilities, Guangdong Province Key Laboratory of Popular High Performance Computers, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, PR China*
[b] *Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, 610051 Chengdu, People's Republic of China*
[c] *URPP Social Networks, Universität Zürich, CH-8050 Switzerland*

Abstract

When investigating the spreading of a piece of information or the diffusion of an innovation, we often lack information on the underlying propagation network. Reconstructing the hidden propagation paths based on the observed diffusion process is a challenging problem which has recently attracted attention from diverse research fields. To address this reconstruction problem, based on static similarity metrics commonly used in the link prediction literature, we introduce new node-node temporal similarity metrics. The new metrics take as input the time-series of multiple independent spreading processes, based on the hypothesis that two nodes are more likely to be connected if they were often infected at similar points in time. This hypothesis is implemented by introducing a time-lag function which penalizes distant infection times. We find that the choice of this time-lag strongly affects the metrics' reconstruction accuracy, depending on the network's clustering coefficient and we provide an extensive comparative analysis of static and temporal similarity metrics for network reconstruction. Our findings shed new light on the notion of similarity between pairs of nodes in complex networks.

Keywords: Information networks, Network reconstruction, Temporal similarity, Innovation diffusion

1. Introduction

Our understanding of social networks is affected by the fact that, typically, we only have incomplete knowledge about the topology of real networks [12, 3]. Aimed at overcoming this shortcoming, the problem of reconstructing missing links has attracted enormous attention from scholars from diverse fields (see [35] for a recent review on the problem). Existing approaches to the network reconstruction problem include the use of local structural metrics [26, 35, 13], global walk-counting methods [22, 35], stochastic block models [18],

Email addresses: manuel.mariani@business.uzh.ch (Manuel Sebastian Mariani), zhoumy2010@gmail.com (Mingyang Zhou)

fitness-based methods [10], structural perturbation analysis [33, 30], machine-learning techniques [17], among many others. Scholars have aimed to identify missing connections in a wide variety of systems, including protein-protein interaction networks [24], neural networks [6], citation networks [11], and social networks [32, 2].

In parallel, there has been recent interest [38, 39, 14, 50, 44, 31] on a different problem of network reconstruction: if we are only provided with information on the outcome of a dynamical process on an unknown propagation network, can we reconstruct the propagation network? The problem – which has been referred to as *latent network reconstruction* [38] – can be included in the broader class of problems that aim to reconstruct the properties of a spreading process (for instance, the seed node [5] or the epidemic parameters [38]) from data on observed realizations of the process. The question is fundamentally different from the traditional link prediction problem [35]: while link prediction studies [35] typically assume that only part of the network is hidden and needs to be reconstructed, here we assume that the topology of the propagation network is completely hidden. The reconstruction problem studied here is important as we often deal with datasets where the propagation network is largely unknown: for instance, the owners of an online e-commerce platform might have complete information on the time-series of users' purchases, but lack information about the social connections between the users which might have affected, to some extent, the observed purchasing patterns.

Existing works have tackled the latent network reconstruction problem from various perspectives. Among the most relevant contributions, Myers et al. [38] addressed the problem through a maximum-likelihood estimation method based on a cascade spreading model, which was further mapped into a convex optimization problem. Gomez-Rodriguez et al. [14] developed a faster maximum-likelihood method based on a cascade propagation model. Shen et al. [44] leveraged compressed sensing theory to map the network reconstruction problem into a convex optimization problem. Such a mapping is non-trivial and model-specific; they solved the problem for the Susceptible-Infected-Susceptible (SIS) and the "contact process" dynamics [44]. The main limitation of these approaches is that they are model-dependent: Different spreading models require the solution of a different set of equations. For example, in the compressed-sensing theory approach, the convex-optimization equations for the SIS and the contact process model differ substantially [44]. Besides, the compressed-sensing approach to network reconstruction can be only applied to sparse networks [44].

On the other hand, other studies [50, 31] have tackled the latent network reconstruction problem by means of simple similarity metrics. With respect to convex optimization [38] and methods based on compressed sensing theory [44], similarity metrics have two main advantages: (1) They do not depend on the specific spreading model considered; (2) Their implementation is faster. Temporal similarity metrics for the latent network reconstruction [31] build on the hypothesis that two nodes are more likely to be connected if independent spreading processes tend to infect them at similar times. A simple way to implement this assumption is to impose, for each pair (i, j) of nodes that are infected by the same spreading process, a

contribution to their similarity s_{ij} in the form of a power-law decreasing function of the time lag between the two infection times [31]. For this reason, we refer to these metrics as temporal similarities with *power-law time lag decay*.

Here, we develop new temporal similarity indexes based on the hypothesis that two nodes are more likely to be connected if independent spreading processes tend to infect them at two consecutive time steps of the dynamics. We refer to the new metrics as temporal similarities with *one-step time lag decay*. Based on the power-law and one-step decay functions, for each of the eight classes of structural similarity metrics considered here, we construct two corresponding temporal similarity metrics. We compare their performance in reconstructing the whole propagation network in both synthetic and real data. By analyzing 40 empirical networks, we provide the first systematic performance comparison of temporal similarity metrics based on different classes of structural similarity metrics.

We find that for the Susceptible-Infected-Recovered (SIR) spreading dynamics [40], for almost all the analyzed networks, the temporal similarities with one-step time lag decay outperform the temporal similarities with a power-law time lag decay. The performance gap is substantially larger for spreading processes sufficiently above their critical point. Besides, among all the classes of similarity metrics considered, we find that the temporal similarity metric with one-step time-lag decay based on the Cosine similarity [43] tends to outperform the other metrics; other competitive classes of similarity are the temporal variants of the Sorensen index [45] and the Jaccard similarity [21]. Results for two additional spreading models (Susceptible-Infected, SI, and Linear Threshold Model, LTM) are in qualitative agreement.

Our findings move the first steps toward an extensive benchmarking of methods for the reconstruction of a hidden topology from the available event time-series of a spreading process. Our work sheds new light on the notion of node-similarity based on the outcome of dynamical processes on networks, and it has potential implications for social network analysis that will be outlined in the Discussion section.

2. Results

2.1. Problem statement

We assume that there is a unipartite network (whose adjacency matrix is denoted by A) whose topology is unknown, and our goal is to reconstruct it. Our available information is the time-stamped list of adoptions of multiple items that diffuse through a given spreading process. Entry $(i, \alpha, t_{i\alpha})$ in this list tells us that node i adopted item α at time $t_{i\alpha}$. The adoption processes considered here is ruled by the SIR dynamics [40]: the "adoption" of item α corresponds to the "infection" during realization α of the SIR dynamics. For this reason, in the following, we will use "adoption" and "infection" interchangeably.

We consider 40 empirical unipartite networks; among these, 20 are information networks (details in the Supplementary Material). We generate the time-series $\{(i, \alpha, t_{i\alpha})\}$ of adoptions by running, for each

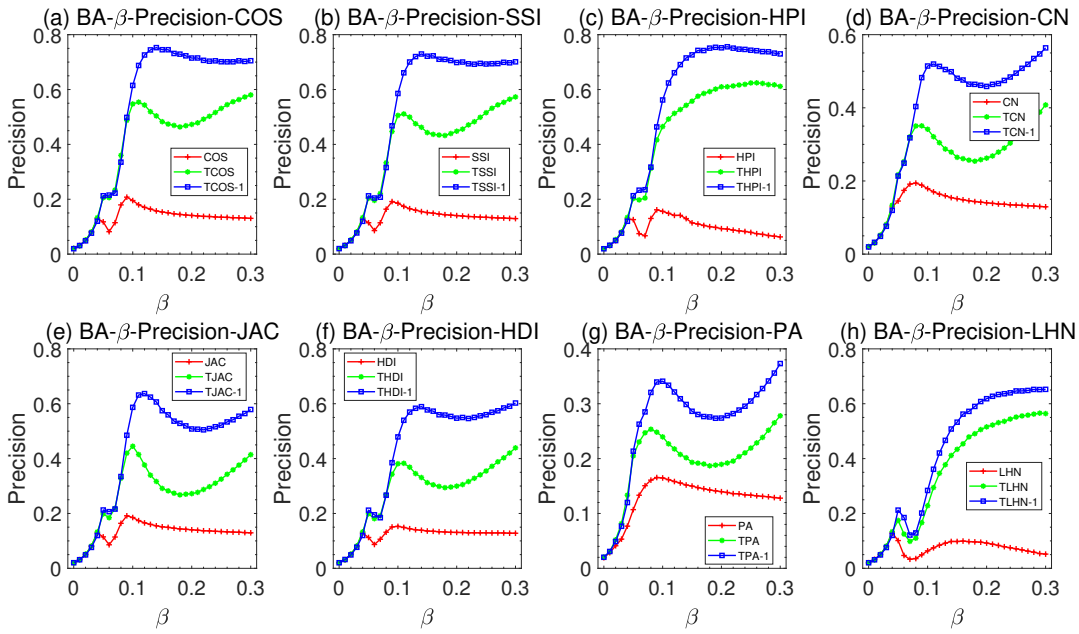

Figure 1: Reconstruction precision of different similarity metrics as a function of β for eight classes of similarity metrics (CN, COS, SSI, HPI, JAC, HDI, PA, LHN), for the SIR dynamics ($f = 0.5$) on BA networks ($N = 500$, $\langle k \rangle = 5$). The results are averaged over 50 independent realizations. For sufficiently large β values, temporal similarity metrics with one-step time-lag decay substantially outperform temporal similarity metrics with power-law time-lag and static metrics.

network, 50 independent realizations of the SIR spreading dynamics initiated by a fraction f of initiators (see Methods for details) [31]. Each independent realization α of the spreading process is therefore interpreted as an item that gradually diffuses across the network. In fact, the time-series $\{(i, \alpha, t_{i\alpha})\}$ can be interpreted as a temporal bipartite network [19]; we denote by R the incidence matrix of the corresponding time-aggregate bipartite network: $R_{i\alpha} = 1$ if node i adopted item α.

We address the following problem. Assuming that we only know $\{i, \alpha, t_{i\alpha}\}$, which is the best method to reconstruct the E edges of A from $\{i, \alpha, t_{i\alpha}\}$? While, in principle, several techniques of network reconstruction can be designed [35, 44, 31], we narrow our focus to similarity metrics that aim to infer the similarity s_{ij} of two nodes i and j based on their co-adoption patterns [50, 31]. The definitions of the metrics of interest are provided in Sections 2.2 and 4.1.

Such similarity metrics produce a ranking of the pairs of nodes (potential edges) in descending order of s_{ij}. Assuming that we know the number of edges E of the underlying propagation network A, the E top-ranked links by s_{ij} form the network $\mathsf{A}^{(s)}$ reconstructed by metric s. It is natural to assess the precision of the metric s_{ij} by measuring the fraction of common links between A and $\mathsf{A}^{(s)}$. This metric is typically referred to as precision in the link prediction [35] and information filtering literature, and we use it to evaluate

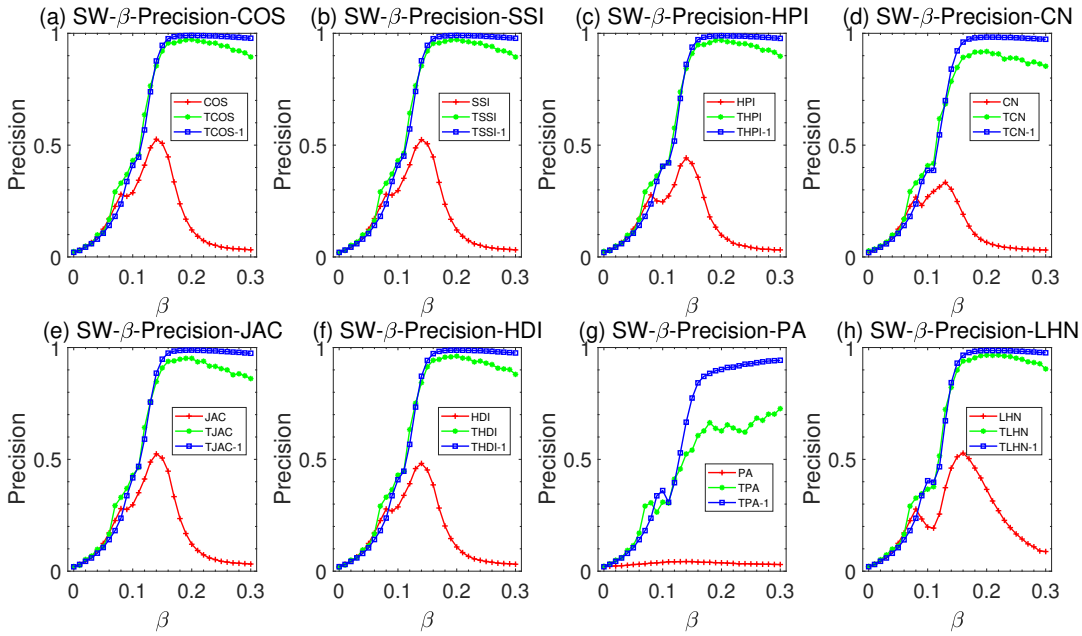

Figure 2: (Color online) Reconstruction precision of different similarity metrics as a function of β for eight classes of similarity metrics (CN, COS, SSI, HPI, JAC, HDI, PA, LHN), for the SIR dynamics ($f = 0.5$) on SW networks ($N = 500$, $P = 0.1$, $\langle k \rangle = 5$). The results are averaged over 50 independent realizations. For sufficiently large β values, temporal similarity metrics with one-step time-lag decay substantially outperform temporal similarity metrics with power-law time-lag and static metrics.

the reconstruction performance of the similarity metrics. The results for another evaluation metric[1] (Area Under the Curve, AUC [34]) are in qualitative agreement with those obtained with the precision (Figs. S8).

2.2. From structural to temporal similarity metrics

We consider here eight classes of structural similarities [35]: common neighbors (CN), Jaccard Index (Jac), Leicht-Holme-Newman Index (LHN), Cosine Index (COS), Sorensen Index (SSI), Hub Promoted Index (HPI), Hub Depressed Index (HDI), Preferential Attachment (PA). These structural metrics have been used by researchers from diverse domains to address various problems in network analysis. They have been applied to the reconstruction of missing links in networks where only a part of the topology is available [12, 34], to the prediction of new connections in social and information systems [32], and to the latent network reconstruction problem studied here as well [50].

For each class[2] X of similarities, we consider the standard static metric [35] (directly denoted as X), and two *temporal* similarity metrics: temporal metrics with the power-law time-lag decay (denoted as TX) [31], and the new temporal metrics with the one-step time-lag decay (denoted as TX1). The last two classes of

[1] Differently from the precision metric, the AUC metric is independent of E.

[2] X is a placeholder here. E.g., X can represent common neighbors CN.

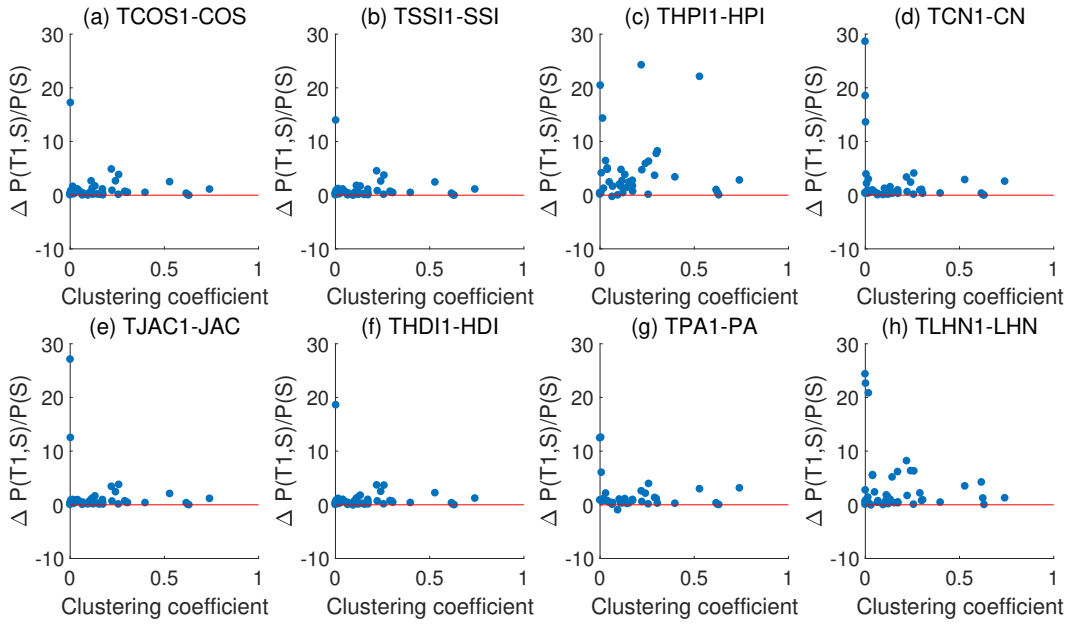

Figure 3: Reconstruction precision relative difference $\Delta P(T1,S)/P(S) = (P(T1) - P(S))/P(S)$ as a function of the network clustering coefficient. Each dot represents an empirical network; we analyzed 40 empirical contact networks. For all classes of similarity, almost all the empirical networks fall above the $P(T1) = P(S)$ red line. We use $\beta = 4\beta_c$ and $f = 0.5$ here. The results are averaged over 50 independent realizations.

metrics differ in how the similarity score of a given pair (i, j) of nodes depends on the *time lag* $t_{i\alpha} - t_{j\alpha}$ between node i's and j's adoption times $t_{i\alpha}$ and $t_{j\alpha}$ for item α. We refer to the Methods section for all the definitions.

To illustrate the main idea behind each class of metrics, we define here the common-neighbors metrics: static common neighbors (CN), temporal common neighbors with a power-law decay of time-lag (TCN), and temporal common neighbors with one-step decay of time lag (TCN1). The common neighbors (CN) of a given pair (i, j) of nodes is simply given by [35]

$$s_{ij}^{CN} = \sum_{\alpha} R_{i\alpha} R_{j\alpha}. \tag{1}$$

According to this definition, two nodes are similar (and, therefore, more likely to be connected in the hidden unipartite network) if they often adopted the same item.

Zeng [50] found that this metric and similar *static* metrics can be used to reconstruct the topology of a hidden network based on the time-series of a spreading dynamics. Subsequently, the static metric proved to be sub-optimal with respect to time-aware metrics [31]. Indeed, while it is plausible that two nodes that often adopt the same item at similar times are more likely to be connected, the same is not necessarily true if the common adoptions happen at very distant points in time: given two adopters i and j, with $t_{i\alpha} \ll t_{j\alpha}$,

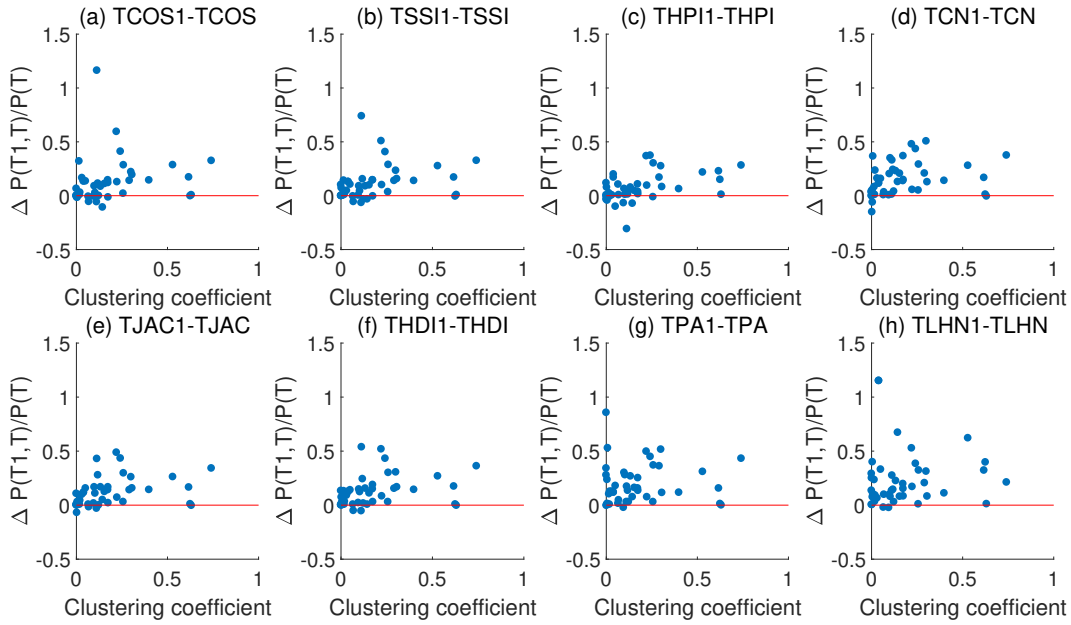

Figure 4: Reconstruction precision relative difference $\Delta P(T1,T)/P(T) = (P(T1) - P(T))/P(T)$ as a function of the network clustering coefficient. Each dot represents an empirical network; we analyzed 40 empirical contact networks. For all classes of similarity, almost all the empirical networks fall above the $P(T1) = P(T)$ red line; the only exceptions are some of the networks with low clustering coefficient. We use $\beta = 4\beta_c$ and $f = 0.5$ here. The results are averaged over 50 independent realizations.

item α might indeed have reached j though a long network path, without the two nodes being directly connected.

To penalize longer time lags, [31] introduced the *temporal common neighbors with power-law time-lag decay* (TCN) as

$$s_{ij}^{TCN} = \sum_{\alpha} R_{i\alpha}\, R_{j\alpha}\, |t_{i\alpha} - t_{j\alpha}|^{-1}(1 - \delta_{t_{i\alpha},t_{j\alpha}}). \tag{2}$$

This time-aware metric significantly outperforms its static counterpart, s^{CN}, in the latent network reconstruction [31]. However, as a consequence of the power-law function, the similarity s^{TCN} of a given pair of nodes receives substantial non-zero contributions also when the two nodes adopt the same item at substantially different times.

In this work, we introduce the *temporal common neighbors with a one-step decay time-lag decay* (TCN1) as

$$s_{ij}^{TCN1} = \sum_{\alpha} R_{i\alpha}\, R_{j\alpha}\, \delta_{|t_{i\alpha}-t_{j\alpha}|,1}. \tag{3}$$

According to this definition, the similarity s^{TCN1} of a given pair (i,j) of nodes only receives a contribution when the two nodes adopt the same item at two consecutive time steps.

Analogous definitions for the other seven classes of similarities and their temporal variants with power-

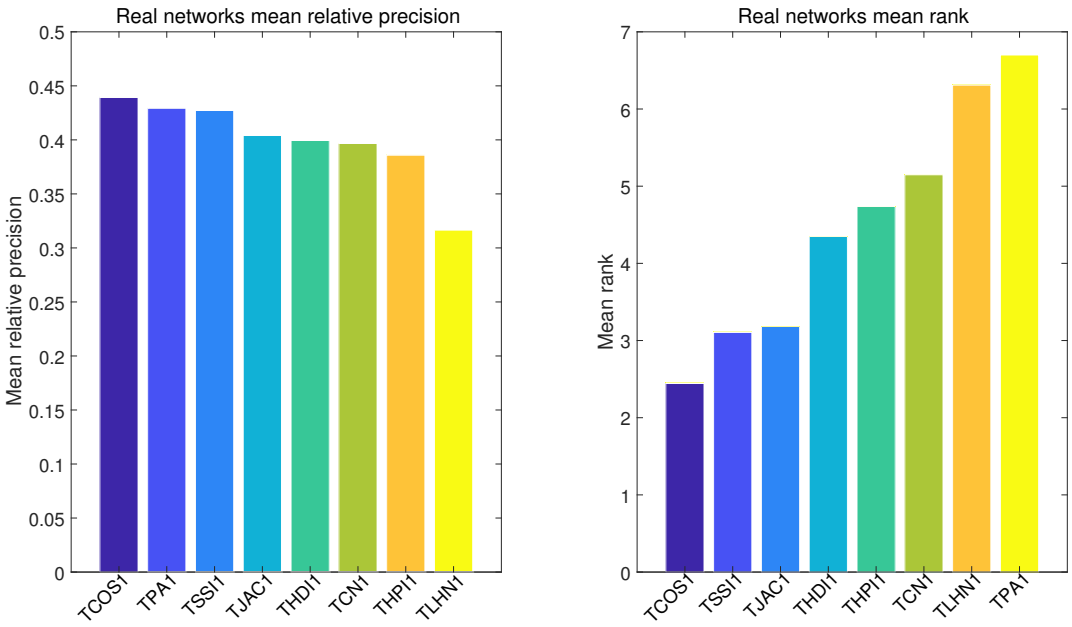

Figure 5: Mean relative precision (higher values correspond to better performance) and mean rank (lower values correspond to better performance) of the eight TX1 metrics. According to both evaluation metrics, TCOS1 is the best-performing metric, followed by TSSI1 and TJAC1. We use $\beta = 4\beta_c$ and $f = 0.5$ here. The results are averaged over 50 independent realizations.

law and one-step time-lag decay are provided in the Methods section. The goal of the rest of the paper is to extensively compare the performance of these metrics in reconstructing both synthetic and empirical networks.

2.3. Reconstruction of synthetic networks

We start our investigation from synthetic networks generated with the Barabási-Albert model [4] (see Methods for the generation details). Fig. 1 shows our reconstruction results: each panel refers to a class of similarities; for each class of similarities (e.g., common neighbors), we show the results for the static metric (CN), the temporal metric with power-law time lag decay (TCN), and the new temporal metric with one-step time lag decay (TCN1). The precision values attained by the metrics are shown as a function of the transmission probability β of the SIR spreading process.

For each considered structural metric (e.g., CN), for sufficiently large β values, the corresponding temporal metric with one-step decay (TCN1) performs significantly better than the corresponding temporal metric with power-law decay (e.g., TCN). As we reduce β, spreading processes tend to die out more rapidly, and it becomes increasingly harder to correctly reconstruct the underlying diffusion network; in the small-β regime, the temporal metrics with a one-step and power-law decay perform similarly. As expected [31], the time-aware metrics significantly outperform the static metric.

Fig. 2 shows analogous results for a small-world network [48] (see Methods for the generation details). We observe again a systematic performance edge of the temporal metrics with a one-step time lag decay over the temporal metrics with power-law time lag decay, yet this gap is smaller than in the BA networks.

2.4. Reconstruction of real networks

Our results on synthetic networks suggest that the temporal metrics with one-step time lag decay reconstruct synthetic contact networks better than the temporal metrics with power-law time lag decay. To further validate this assertion, we analyzed 40 empirical contact networks of diverse nature including 20 information networks (details in the Supplementary Material).

For almost all the analyzed datasets, the temporal metrics with one-step time lag decay substantially improve the reconstruction accuracy with respect to both static (Fig. 3) and temporal metrics with power-law time lag decay (Fig. 4). The only networks where the temporal metrics with power-law time-lag decay can outperform the temporal metrics with one-step time-lag decay are those with low clustering coefficient[3]. This is intuitive: In a network with lower clustering, it is less likely that two non-connected nodes are reached by long propagation paths. This mitigates the advantage of considering only adoptions with one-step time lag when computing the similarity score of a given pair of nodes.

The results in Fig. 4 were obtained with $\beta = 4\,\beta_c$, where β_c is the epidemic threshold [40]. As expected from the synthetic network analysis, we find that for larger β values (Fig. S1), the one-step time lag metrics show better reconstruction accuracy for the vast majority of datasets and considered metrics. On the other hand, for lower β values, there is not a clear advantage of the metrics with the one-step time-lag decay (Figs. S2-S3).

So far, we have compared similarities of the same class (e.g., common neighbors) with different time-lag decay functions. A natural question arises: what is the relative performance of the eight temporal metrics TX1 with one-step time-lag decay obtained from the eight different classes of similarities? We compare the eight metrics' performance across the 40 empirical datasets considered here. We refer to Figs. S4-S5 for the results on individual datasets. To gain a general understanding of the metrics' performance, we aggregate the metrics' performance over the analyzed networks. To this end, we consider two evaluation metrics: the metrics' mean rank [37] and the mean relative precision.

To compute the metrics' mean rank, for each dataset d, we rank the eight TX1 metrics in order of decreasing precision. We denote by $r_d(s) \in \{1, 2, \ldots, 8\}$ the ranking position metric s for dataset d. Given D analyzed empirical networks ($D = 40$ in our work), the mean rank $\overline{r(s)}$ of metric s is simply defined as

[3] In our work, we use the average local clustering coefficient as a metric for clustering. For each node i in the network, we calculate the number K_i of existing edges that connect nodes that are connected with i, and the maximum number E_i of possible links between i's neighbors. For an undirected graph, $E_i = k_i\,(k_i - 1)/2$. Finally, we define i's local clustering coefficient $C_i = K_i/E_i$, and the network's clustering coefficient as $C = N^{-1}\sum_i C_i$.

$\overline{r(s)} = D^{-1}\sum_{d=1}^{D} r_d(s)$. Better performing metric should exhibit lower mean rank values [37]. In addition, denoting by $P_d(s)$ the precision achieved by metric s in dataset d, we define the mean relative precision $\overline{P(s)}$ of metric s as $D^{-1}\sum_{d=1}^{D} P_d(s)/\max_{s'}\{P_d(s')\}$. Better performing metric should exhibit larger mean relative precision values.

Both evaluation metrics lead to the same overall conclusion: on average, the TCOS1 (temporal Cosine with one-step time-lag decay) metric

$$s_{ij}^{TCOS1} = \frac{\sum_\alpha R_{i\alpha}R_{j\alpha}\delta_{|t_{i\alpha}-t_{j\alpha}|,1}}{\sqrt{\sum_\alpha R_{i\alpha}\sum_\alpha R_{j\alpha}}} \tag{4}$$

is the best-performing metric, followed by TSS1 and TJAC1 (see Methods for their definition). While TCOS1 provides us with a computationally fast metric to reconstruct the hidden topology, its mean precision is 0.349. This leaves the door open for future performance improvements, possibly based on new similarity metrics or more sophisticated methods.

3. Discussion

Our work provided a systematic benchmarking of temporal similarity metrics with respect to their accuracy in reconstructing a hidden network topology. The reconstruction was more accurate for SIR spreading processes with a large transmission probability, i.e., in the supercritical regime. On both real and synthetic networks, we found that temporal metrics with one-step time-lag decay perform systematically better than metrics with power-law time-lag decay. Besides, we found that the temporal cosine metric with one-step time-lag decay is the best-performing metric. Differently from maximum-likelihood methods [14] and compressed-sensing theory approaches [38], the temporal similarity metrics considered here are general and not restricted to a specific dynamics. In this sense, they can be interpreted not only as parsimonious and effective reconstruction tools, but also as general baselines against which more sophisticated, model-specific reconstruction techniques can be evaluated. While we focused on the SIR dynamics, we also assessed the metrics' performance for two additional spreading models: the Susceptible-Infected (SI) model [1, 46] and the Linear Threshold Model (LTM) [16, 23, 8]. The results obtained for these two models are in qualitative agreement with the results obtained for the SIR model (Figs. S6-S7), supporting the generality of our conclusions.

Our study paves the way for several extensions. Temporal similarity metrics might be applied to other network reconstruction problems, such as the problem where part of the topology is known [34] and the matching of user accounts across different domains or devices [7, 29]. Even more intriguingly, one can attempt to reconstruct the hidden topology of a social network based on the observed dynamics of real diffusion processes. For instance, from the observed spreading dynamics of many pieces of information, one might attempt to reconstruct propagation networks in social media [41] and e-commerce platforms [36].

The results presented here support metrics based on one-step time lags as the best-performing ones in the latent network reconstruction task. While the time step of the dynamics is unambiguously defined for simulated processes, the same does not hold for real spreading processes. Using temporal similarity metrics to reconstruct propagation topologies based on real time-series data will likely require us to first identify the typical timescale needed for a given piece of information to be transmitted from an individual to another, and then to use this typical timescale as the time-lag parameter in the similarity metric.

Finally, our work contributes to the rich literature on similarity on social and information networks [28, 27, 47, 15, 9, 42]. Previous research has stressed the role of structural similarity metrics, i.e., similarity metrics based either on the time-aggregate contact network of individuals (who is connected to whom) [20, 49, 25] or on the time-aggregate user-item bipartite adoption network (who collected what) [27, 50]. Here, we combined structure and temporal information (who collected what at which time) to define temporal similarity metrics that are effective in the propagation network task. We envision that future research on social and information network analysis might further develop simple yet well-performing time-aware metrics for network reconstruction.

4. Methods

4.1. Temporal similarity metrics

For each class C of similarity metrics, we define three metrics: a static metric C, a temporal metric with power-law time-lag decay TC, and a temporal metric with one-step time-lag decay $TC1$. In our work, we consider eight classes C of similarities: Common Neighbors (CN), Jaccard (Jac), Cosine (COS), Leicht-Holme-Newman (LHN), Sorensen Index (SSI), Hub-promoted Index (HPI), Preferential Attachment (PA), Hub-depressed Index (HDI). As we already defined the three CN similarities in the main text, we define here the metrics based on the seven additional classes.

Jaccard (Jac) similarity. We define three metrics:

- Jaccard similarity (Jac):

$$s_{ij}^{Jac} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha}}{\sum_\alpha (R_{i\alpha} + R_{j\alpha} - R_{i\alpha} R_{j\alpha})}. \tag{5}$$

- Temporal Jaccard similarity with power-law time-lag decay (TJac):

$$s_{ij}^{TJac} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha} |t_{i\alpha} - t_{j\alpha}|^{-1} (1 - \delta_{t_{i\alpha}, t_{j\alpha}})}{\sum_\alpha (R_{i\alpha} + R_{j\alpha} - R_{i\alpha} R_{j\alpha})}. \tag{6}$$

- Temporal Jaccard similarity with one-step time-lag decay (TJac1):

$$s_{ij}^{TJac1} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha} \delta_{|t_{i\alpha} - t_{j\alpha}|, 1}}{\sum_\alpha (R_{i\alpha} + R_{j\alpha} - R_{i\alpha} R_{j\alpha})}. \tag{7}$$

Cosine (COS) similarity. We define three metrics:

- Cosine similarity (COS):

$$s_{ij}^{COS} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha}}{\sqrt{\sum_\alpha R_{i\alpha} \sum_\alpha R_{j\alpha}}}. \tag{8}$$

- Temporal Cosine similarity with power-law time-lag decay (TCOS):

$$s_{ij}^{TCOS} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha} |t_{i\alpha} - t_{j\alpha}|^{-1} (1 - \delta_{t_{i\alpha}, t_{j\alpha}})}{\sqrt{\sum_\alpha R_{i\alpha} \sum_\alpha R_{j\alpha}}}. \tag{9}$$

- Temporal Cosine similarity with one-step time-lag decay (TCOS1):

$$s_{ij}^{TCOS1} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha} \delta_{|t_{i\alpha} - t_{j\alpha}|, 1}}{\sqrt{\sum_\alpha R_{i\alpha} \sum_\alpha R_{j\alpha}}}. \tag{10}$$

Leicht-Holme-Newman Index (LHN) similarity. We define three metrics:

- Leicht-Holme-Newman Index similarity (LHN):

$$s_{ij}^{LHN} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha}}{\sum_\alpha R_{i\alpha} \sum_\alpha R_{j\alpha}}. \tag{11}$$

- Temporal Leicht-Holme-Newman Index similarity with power-law time-lag decay (TLHN):

$$s_{ij}^{TLHN} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha} |t_{i\alpha} - t_{j\alpha}|^{-1} (1 - \delta_{t_{i\alpha}, t_{j\alpha}})}{\sum_\alpha R_{i\alpha} \sum_\alpha R_{j\alpha}}. \tag{12}$$

- Temporal Leicht-Holme-Newman Index similarity with one-step time-lag decay (TLHN1):

$$s_{ij}^{TLHN1} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha} \delta_{|t_{i\alpha} - t_{j\alpha}|, 1}}{\sum_\alpha R_{i\alpha} \sum_\alpha R_{j\alpha}}. \tag{13}$$

Sørensen Index (SSI) similarity. We define three metrics:

- Sørensen Index similarity (SSI):

$$s_{ij}^{SSI} = \frac{2 \times \sum_\alpha R_{i\alpha} R_{j\alpha}}{\sum_\alpha R_{i\alpha} + \sum_\alpha R_{j\alpha}}. \tag{14}$$

- Temporal Sørensen Index similarity with power-law time-lag decay (TSSI):

$$s_{ij}^{TSSI} = \frac{2 \times \sum_\alpha R_{i\alpha} R_{j\alpha} |t_{i\alpha} - t_{j\alpha}|^{-1} (1 - \delta_{t_{i\alpha}, t_{j\alpha}})}{\sum_\alpha R_{i\alpha} + \sum_\alpha R_{j\alpha}}. \tag{15}$$

- Temporal Sørensen Index similarity with one-step time-lag decay (TSSI1):

$$s_{ij}^{TSSI1} = \frac{2 \times \sum_\alpha R_{i\alpha} R_{j\alpha} \delta_{|t_{i\alpha} - t_{j\alpha}|, 1}}{\sum_\alpha R_{i\alpha} + \sum_\alpha R_{j\alpha}}. \tag{16}$$

Hub Promoted Index (HPI) similarity. We define three metrics:

- Hub Promoted Index similarity (HPI):

$$s_{ij}^{HPI} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha}}{min\{\sum_\alpha R_{i\alpha}, \sum_\alpha R_{j\alpha}\}} \tag{17}$$

- Temporal Hub Promoted Index similarity with power-law time-lag decay (THPI):

$$s_{ij}^{THPI} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha} |t_{i\alpha} - t_{j\alpha}|^{-1} (1 - \delta_{t_{i\alpha}, t_{j\alpha}})}{min\{\sum_\alpha R_{i\alpha}, \sum_\alpha R_{j\alpha}\}} \tag{18}$$

- Temporal Hub Promoted Index similarity with one-step time-lag decay (THPI1):

$$s_{ij}^{THPI1} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha} \delta_{|t_{i\alpha} - t_{j\alpha}|, 1}}{min\{\sum_\alpha R_{i\alpha}, \sum_\alpha R_{j\alpha}\}} \tag{19}$$

Hub Depressed Index (HDI) similarity. We define three metrics:

- Hub Depressed Index similarity (HDI):

$$s_{ij}^{HDI} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha}}{max\{\sum_\alpha R_{i\alpha}, \sum_\alpha R_{j\alpha}\}} \tag{20}$$

- Temporal Hub Depressed Index similarity with power-law time-lag decay (THDI):

$$s_{ij}^{THDI} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha} |t_{i\alpha} - t_{j\alpha}|^{-1} (1 - \delta_{t_{i\alpha}, t_{j\alpha}})}{max\{\sum_\alpha R_{i\alpha}, \sum_\alpha R_{j\alpha}\}} \tag{21}$$

- Temporal Hub Depressed Index similarity with one-step time-lag decay (THDI1):

$$s_{ij}^{THDI1} = \frac{\sum_\alpha R_{i\alpha} R_{j\alpha} \delta_{|t_{i\alpha} - t_{j\alpha}|, 1}}{max\{\sum_\alpha R_{i\alpha}, \sum_\alpha R_{j\alpha}\}} \tag{22}$$

Preferential Attachment (PA) similarity. We define three metrics:

- Preferential Attachment similarity (PA):

$$s_{ij}^{PA} = \sum_\alpha R_{i\alpha} \sum_\alpha R_{j\alpha}. \tag{23}$$

- Temporal Preferential Attachment similarity with power-law time-lag decay (TPA):

$$s_{ij}^{TPA} = \sum_\alpha R_{i\alpha} \sum_\alpha R_{j\alpha} |t_{i\alpha} - t_{j\alpha}|^{-1} (1 - \delta_{t_{i\alpha}, t_{j\alpha}}). \tag{24}$$

- Temporal Preferential Attachment similarity with one-step time-lag decay (TPA1):

$$s_{ij}^{TPA1} = \sum_\alpha R_{i\alpha} \sum_\alpha R_{j\alpha} \delta_{|t_{i\alpha} - t_{j\alpha}|, 1}. \tag{25}$$

In all the temporal similarity methods above, we set $(t_{i\alpha} - t_{j\alpha})^{-1} = 0$ when $t_{i\alpha} = t_{j\alpha}$. Note that in the TC metrics, the factor $1 - \delta_{t_{i\alpha}, t_{j\alpha}}$ makes sure that events where $t_{i\alpha} = t_{j\alpha}$ do not contribute to the similarity. Indeed, when $t_{i\alpha} = t_{j\alpha}$, i is not the node that infected j; therefore, i and j are unlikely to be connected in the networks. Note that in other problems such as link prediction and recommendation, the case $t_{i\alpha} = t_{j\alpha}$ may need to be treated differently.

4.2. SIR spreading dynamics

In the SIR model, each node is in one of the three states: Susceptible (S), Infected (I), Recovered (R). Each node has a probability f to be an initiator of the spreading process; therefore, there are $f \times N$ simultaneous initiators, on average, for each spreading process. At each time step, each infected node can infect each of its neighbors with probability β; each infected node can recover with probability μ. For simplicity, we fix $\mu = 1$ (each node recovers one step after having been infected). The process ends when there are no more infected nodes in the system. For each empirical network, we run 50 independent realizations of the SIR dynamics. For each process α, we record the temporal list of the nodes infected by that process. The bipartite adjacency matrix R records which nodes were infected by which process: $R_{i\alpha} = 1$ if i has been infected by α, whereas $R_{i\alpha} = 0$ otherwise. If $R_{i\alpha} = 1$, the time step at which i was infected by α is recorded in $t_{i\alpha}$.

4.3. Generation of the synthetic networks

We use two well-known models for the generation of synthetic networks: the Barabási-Albert (BA) model [4], and the Small-World (SW) model [48].

Barabási-Albert (BA). We generate networks composed of $N = 500$ nodes. Our initial condition is a regular network where each node composed of $m_0 = 9$ nodes; each initial node has the degree equal to $\langle k \rangle = 5$. At each time step t, we add a new node to the network. The new node connects with $\langle k \rangle$ preexisting nodes; the probability that a preexisting node i is selected is proportional to its degree $k_i(t)$ at time t.

Small-World (SW). We start from a regular ring lattice composed of $N = 500$ nodes and degree $k = \langle k \rangle = 5$: we connect each of the N nodes with its nearest k neighbors. We rewire each link with probability p – in this work, we set $p = 0.1$. More specifically, for each node i, we select a node j from its neighbors and we extract a random number r from the uniform distribution in $(0, 1)$. If p is larger than r, we and remove the edge between node i and node j, we randomly select a node m, and we establish an edge between node i and node m.

Competing interests

The authors declare that they have no competing interests.

Author's contribution

The work presented in this paper corresponds to a collaborative development by all authors. Conceptualization, H.L., M.S.M, and M-Y.Z.; Data Curation, M-K.L. and H.L.; Formal Analysis, H.L., M-K.L. and M.S.M.; Funding Acquisition, H.L. and M-Y.Z.; Methodology, M.S.M.; Resources, H.L. and M-Y.Z.; Software, M-K.L. and X-T.W.; Writing—Original Draft, M.S.M., M-K.L., M-Y.Z., X-T.W. and H.L.

Acknowledgements

We wish to thank Prof. Ginestra Bianconi and Prof. Chi Ho Yeung for providing us valuable suggestions. H.L and M.Y.Z acknowledge financial support from the National Natural Science Foundation of China (Grant Nos. 61803266, 61703281), Guangdong Province Natural Science Foundation (Grant Nos. 2016A030310051,2017A030310374, 2017B030314073), Guangdong Pre-national project (Grant Nos. 2014GK XM054), Shenzhen Fundamental Research Foundation (JCYJ20160520162743717, JCYJ20150529164-656096), Natural Science Foundation of SZU (Grant No. 2016-24), Foundation for Distinguished Young Talents in Higher Education of Guangdong, China(Grant No. 2015K-QNCX143). MSM acknowledges the University of Zurich for support through the URPP Social Networks.

References

[1] R. M. Anderson, R. M. May, Infectious diseases of humans: dynamics and control, Oxford university press, Oxford, England, UK, 1992.

[2] L. Backstrom, J. Leskovec, Supervised random walks: predicting and recommending links in social networks, in: Proceedings of the 4th ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 635–644.

[3] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The role of social networks in information diffusion, in: Proceedings of the 21st International Conference on World Wide Web, ACM, 2012, pp. 519–528.

[4] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.

[5] D. Brockmann, D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, Science 342 (6164) (2013) 1337–1342.

[6] C. V. Cannistraci, G. Alanis-Lobato, T. Ravasi, From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks, Sci. Rep. 3 (2013) 1613.

[7] W. Chen, H. Yin, W. Wang, L. Zhao, X. Zhou, Effective and efficient user account linkage across location based social networks, in: Proceedings of the 34th IEEE International Conference on Data Engineering, IEEE, 2018, pp. 1085–1096.

[8] W. Chen, Y. Yuan, L. Zhang, Scalable influence maximization in social networks under the linear threshold model, in: Proceedings of the 10th IEEE International Conference on Data Mining, IEEE, 2010, pp. 88–97.

[9] Y. Chen, N. Crespi, A. M. Ortiz, L. Shu, Reality mining: A prediction algorithm for disease dynamics based on mobile big data, Inf. Sci. 379 (2017) 82–93.

[10] G. Cimini, T. Squartini, D. Garlaschelli, A. Gabrielli, Systemic risk analysis on reconstructed economic and financial networks, Sci. Rep. 5 (2015) 15758.

[11] V. Ciotti, M. Bonaventura, V. Nicosia, P. Panzarasa, V. Latora, Homophily and missing links in citation networks, EPJ Data Sci. 5 (1) (2016) 7.

[12] A. Clauset, C. Moore, M. E. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (7191) (2008) 98.

[13] S. Daminelli, J. M. Thomas, C. Durán, C. V. Cannistraci, Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks, New J. Phys. 17 (11) (2015) 113037.

[14] M. Gomez-Rodriguez, J. Leskovec, A. Krause, Inferring networks of diffusion and influence, ACM Trans. Knowl. Discov. Data 5 (4) (2012) 21.

[15] M. Gong, J. Yan, B. Shen, L. Ma, Q. Cai, Influence maximization in social networks based on discrete particle swarm optimization, Inf. Sci. 367 (2016) 600–614.

[16] M. Granovetter, Threshold models of collective behavior, Am. J. Sociol. 83 (6) (1978) 1420–1443.

[17] A. Grover, J. Leskovec, Node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 855–864.

[18] R. Guimerà, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proc. Natl. Acad. Sci. 106 (52) (2009) 22073–22078.

[19] P. Holme, J. Saramäki, Temporal networks, Phys. Rep. 519 (3) (2012) 97–125.

[20] Z. Huang, D. K. Lin, The time-series link prediction problem with applications in communication surveillance, INFORMS J. Comput. 21 (2) (2009) 286–303.

[21] P. Jaccard, Étude comparative de la distribution florale dans une portion des alpes et des jura, Bull. Soc. Vaudoise Sci. Nat. 37 (1901) 547–579.

[22] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.

[23] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003, pp. 137–146.

[24] I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D.-K. Kim, N. Kishore, T. Hao, Network-based prediction of protein interactions, bioRxiv (2018) 275529.

[25] D. H. Lee, P. Brusilovsky, How to measure information similarity in online social networks: A case study of citeulike, Inf. Sci. 418 (2017) 46–60.

[26] E. A. Leicht, P. Holme, M. E. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2) (2006) 026120.

[27] R.-H. Li, J. Xu Yu, X. Huang, H. Cheng, Robust reputation-based ranking on bipartite rating networks, in: Proceedings of the 2012 SIAM International Conference on Data Mining, SIAM, 2012, pp. 612–623.

[28] R.-H. Li, J. X. Yu, J. Liu, Link prediction: the power of maximal entropy random walk, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, 2011, pp. 1147–1156.

[29] Y. Li, Z. Zhang, Y. Peng, H. Yin, Q. Xu, Matching user accounts based on user generated content across social networks, Future Gener. Comput. Syst. 83 (2018) 104–115.

[30] H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, M.-Y. Zhou, Ranking in evolving complex networks, Phys. Rep. 689 (2017) 1–54.

[31] H. Liao, A. Zeng, Reconstructing propagation networks with temporal similarity, Sci. Rep. 5 (2015) 11404.

[32] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inform. Sci. Tech. 58 (7) (2007) 1019–1031.

[33] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, H. E. Stanley, Toward link predictability of complex networks, Proc. Natl. Acad. Sci. 112 (8) (2015) 2325–2330.

[34] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (6) (2011) 1150–1170.

[35] V. Martínez, F. Berzal, J.-C. Cubero, A survey of link prediction in complex networks, ACM Comput. Surv. 49 (4) (2017) 69.

[36] M. Medo, M. S. Mariani, A. Zeng, Y.-C. Zhang, Identification and impact of discoverers in online social systems, Sci. Rep. 6 (2016) 34218.

[37] A. Muscoloni, C. V. Cannistraci, Local-ring network automata and the impact of hyperbolic geometry in complex network link-prediction, 2017, arXiv preprint arXiv:1707.09496 .

[38] S. Myers, J. Leskovec, On the convexity of latent social network inference, in: Advances in Neural Information Processing Systems 23, 2010, pp. 1741–1749.

[39] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguná, D. Krioukov, Popularity versus similarity in growing networks, Nature 489 (7417) (2012) 537.

[40] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Mod.

Phys. 87 (3) (2015) 925.

[41] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, H. A. Makse, Searching for superspreaders of information in real-world social media, Sci. Rep. 4 (2014) 5547.

[42] S. Peng, A. Yang, L. Cao, S. Yu, D. Xie, Social influence modeling using information theory in mobile social networks, Inf. Sci. 379 (2017) 146–159.

[43] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, USA, 1986.

[44] Z. Shen, W.-X. Wang, Y. Fan, Z. Di, Y.-C. Lai, Reconstructing propagation networks with natural diversity and identifying hidden sources, Nat. Commun. 5 (2014) 4323.

[45] T. Sørensen, A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons, Biologiske skrifter, I kommission hos E. Munksgaard, Argentina, Schleswig, Sun, 1948.

[46] M. Starnini, A. Machens, C. Cattuto, A. Barrat, R. Pastor-Satorras, Immunization strategies for epidemic processes in time-varying contact networks, J. Theor. Biol. 337 (2013) 89–100.

[47] Z. Sun, Q. Peng, J. Lv, J. Zhang, A prediction model of post subjects based on information lifecycle in forum, Inf. Sci. 337 (2016) 59–71.

[48] D. J. Watts, S. H. Strogatz, Collective dynamics of small-worldnetworks, Nature 393 (6684) (1998) 440.

[49] P. Xia, L. Zhang, F. Li, Learning similarity with cosine similarity ensemble, Inf. Sci. 307 (2015) 39–52.

[50] A. Zeng, Inferring network topology via the propagation process, J. Stat. Mech. Theory Exp. 2013 (11) (2013) 11010.